Inform : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi
Vol.11 No.1, January 2026, P-ISSN : 2502-3470, E-ISSN : 2581-0367

Performance Analysis of Support Vector Machine with Hyperparameter Tuning for Engagement Rate Classification of TikTok Digital Marketing Content

Muhammad Dzar Algifahri[1], Sriani[2]

[1,2]*Computer Science Department, Universitas Islam Negeri Sumatera Utara Medan, Indonesia*

[1]algidzarfahri07@gmail.com (*)

[2]sriani@uinsu.ac.id

Received: 2025-12-17; Accepted: 2026-01-29; Published: 2026-02-02

Abstract— The effectiveness of TikTok digital marketing content often varies, creating challenges for marketers in achieving consistent audience engagement. Differences in posting time, upload frequency, video length, and music selection make it difficult to distinguish engagement rate (ER) levels accurately. This paper investigates key factors affecting engagement and proposes a classification framework to separate high- and low-engagement TikTok content. A machine learning approach using the Support Vector Machine (SVM) algorithm is applied to a dataset of TikTok videos collected between 2023 and 2025. From an initial dataset of 2,992 videos, 1,638 representative samples were retained after data cleaning and creator-level filtering. The research process involves feature engineering, engagement-based labelling with a 5% ER threshold, data normalisation, and dataset partitioning with an 80:20 training–testing split. The baseline SVM model achieved an accuracy of 70.43%, indicating limited ability to distinguish low-engagement content. After systematic hyperparameter tuning, the optimised linear SVM model demonstrated improved performance, achieving an accuracy of 88.41% with an optimal regularisation parameter (C = 100) and more balanced classification results. Model interpretation indicates that video duration, temporal attributes, and audio characteristics play important roles in separating engagement levels. The proposed framework is intended for post-hoc engagement classification rather than engagement prediction, providing interpretable insights to support TikTok digital marketing strategy optimization.

Keywords— TikTok; Digital Marketing; Engagement Rate; Data Mining; Support Vector Machine.

I. INTRODUCTION

Engagement Rate (ER) is a metric that measures audience interaction with social media content and reflects how well a social media account engages its followers [1]. Rapid advancements in information technology and the widespread adoption of the internet over the past decade have fundamentally reshaped patterns of communication, information access, and business activities [2]. Within this digital ecosystem, TikTok has emerged as one of the fastest-growing social media platforms, driven by its short-form video format and algorithm-based content distribution. As of January 2023, TikTok recorded approximately 1.05 billion active users worldwide, with Indonesia ranking among the largest user bases at nearly 109.90 million users, positioning the platform as a dominant medium for digital interaction and content consumption [3]. The availability of creative tools such as visual effects, filters, and simplified video production mechanisms further enhances user participation and interaction, reinforcing TikTok's role as a high-engagement social media platform [4].

From a digital marketing perspective, TikTok provides a distinctive environment that combines engaging audiovisual content, personalized recommendation systems, and integrated commercial features such as TikTok Shop, affiliate programs, and paid advertising. Empirical studies conducted in Indonesia demonstrate that TikTok is particularly effective in increasing brand visibility and driving sales performance, especially for micro, small, and medium enterprises (MSMEs) and fashion-related businesses [5]. These findings align with existing studies indicating that short-form video content distributed on social media platforms is effective at attracting audience attention, stimulating engagement, and driving conversion-oriented outcomes, including increased store visits and sales [6].

Engagement rate is widely used as a metric to quantify how actively audiences interact with social media content and is commonly derived from user interaction indicators such as likes, comments, and shares, which are effective predictors in machine learning–based engagement modelling [7]. Engagement rate is operationalised as the proportion of cumulative user interactions (likes, comments, shares, and saves) relative to total video views. This formulation allows engagement rate to function not only as a descriptive performance indicator but also as a target variable for binary classification, representing the intensity of audience engagement. Industry benchmarks suggest that engagement rates between 1% and 5% indicate satisfactory performance, while values exceeding 5% reflect strong to very strong audience engagement[8][9]. In addition, multiple social media analytics sources have reported that typical engagement rates on TikTok average around 4–5%, providing empirical support for using a 5% threshold to distinguish high-engagement content. Based on these widely adopted benchmarks and empirical observations, this study uses a 5% threshold to distinguish between low- and high-engagement content. Selecting this threshold provides a standardised, non-arbitrary

DOI : https://doi.org/10.25139/inform.v11i1.11362

criterion that enhances comparability across studies and enables consistent binary labelling of TikTok content for engagement rate classification [10].

Previous studies have identified multiple content-related and temporal factors that significantly influence engagement levels on TikTok [11]. Variables such as video duration, upload time, day of publication, use of trending or original music, and behavioural differences between weekday and weekend audiences directly affect user interaction patterns. Consequently, modelling these variables is essential for understanding engagement dynamics and for developing predictive mechanisms that support content optimisation strategies [12]. However, the increasing volume and complexity of TikTok data render manual analysis inefficient and impractical, necessitating the use of data mining and machine learning approaches.

Support Vector Machine (SVM) is frequently used in social media analytics as a classification technique that separates data into distinct categories, particularly when dealing with datasets with complex feature representations. Several empirical studies have reported that SVM delivers stable, reliable performance in classification tasks, including sentiment and opinion analysis on social media platforms [13]. In the Indonesian context, SVM achieved competitive classification accuracy compared with Logistic Regression and Naïve Bayes in analysing TikTok user reviews. However, their study primarily focused on sentiment classification and did not specifically address engagement rate analysis or digital marketing applications [14]. Similarly, the research [15] found that engagement with TikTok video content was significantly influenced by various content-related factors, such as visual quality, the use of trending music, and content distribution strategies, highlighting the importance of content attributes in shaping audience engagement. In addition, SVM was selected to prioritise interpretability and feature-level explanations, which are essential for translating model outputs into actionable digital marketing insights, rather than solely maximising predictive performance [15].

Compared to more complex models such as deep learning architectures, SVM offers advantages in interpretability, computational efficiency, and suitability for datasets with limited sample sizes, making it particularly appropriate for engagement rate classification tasks that require both predictive accuracy and explanatory insights [16]. Moreover, previous studies in social media analytics have demonstrated that systematic hyperparameter tuning can significantly enhance SVM performance in classification tasks, especially on large-scale, imbalanced social media datasets [17]. However, despite the growing body of research on TikTok engagement and the use of SVM in social media analytics, existing studies have largely focused on sentiment analysis or descriptive engagement factors, with limited attention to SVM-based classification of optimised engagement rates in digital marketing contexts. In particular, the integration of systematic hyperparameter tuning with engagement rate classification and the translation of model outputs into actionable marketing insights remain underexplored. Addressing this research gap, the present study proposes an optimized SVM-based framework for engagement rate classification on TikTok digital marketing content, aiming to deliver both robust predictive performance and practically relevant strategic insights. The main contributions of this study include the application of an optimized SVM framework for engagement rate classification, the integration of systematic hyperparameter tuning, and the translation of model outputs into practical upload and content optimization strategies for TikTok digital marketing.

II. Research Methodology

This section outlines the research approach for categorising TikTok content related to digital marketing by analysing engagement levels using the Support Vector Machine (SVM) technique. The methodological structure involves a series of consecutive steps, including gathering data, preparing it for analysis, developing features, assigning labels, applying the model, and assessing its effectiveness. The complete sequence of the study's workflow is depicted in Fig.1, which shows the organised progression followed in this investigation. The analysis begins with data acquisition, followed by pre-processing and feature construction, and then proceeds to SVM-based classification and evaluation. This structured workflow ensures that each analytical stage contributes to improving data quality and model reliability.

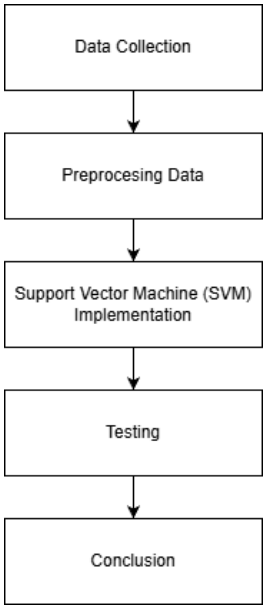

Fig.1. Research Design Workflow

A. Data Collection

TikTok digital marketing data were collected through web scraping of publicly accessible TikTok accounts over the 2023–2025 period. Data collection was conducted using the Apify web scraping platform to retrieve publicly accessible TikTok content. The scraping process complied with ethical research standards by collecting only public data and excluding any private or restricted user information.

The data acquisition process focused on video-level information, including user interaction metrics (likes, comments, shares, saves, and views) and content-related

DOI : https://doi.org/10.25139/inform.v11i1.11362

attributes (video duration, upload time, and music characteristics). All collected records were stored in CSV format to support subsequent data processing and analysis. The dataset attributes used in this research are summarised in Table I, which presents the interaction metrics and content metadata used to model TikTok engagement rate patterns. Data labelled as 2025 correspond to videos obtained via real-time web scraping conducted in early 2025. For transparency and reproducibility, the dataset used in this study is publicly available at https://github.com/Algifhri31/Dataset_Journal.

TABLE I
DATASET ATTRIBUTES DESCRIPTION

Column Name	Description
authorMeta.name	Creator or account owner name
text	Video caption or description
diggCount	Number of likes
shareCount	Number of shares
playCount	Number of video views
commentCount	Number of comments
collectCount	Number of saves
videoMeta.duration	Video duration (seconds)
musicMeta.musicName	Music title used in the video
musicMeta.musicAuthor	Music creator or artist
musicMeta.musicOriginal	Original audio status
createTimeISO	Upload time in ISO format.
webVideoUrl	TikTok video URL

B. Data Pre-processing

Data pre-processing was performed to ensure analytical quality, consistency, and reliability before feature engineering and model development. This stage plays an essential role in preparing TikTok data for machine learning–based classification by reducing noise, redundancy, and potential bias commonly present in raw social media data. The pre-processing process includes data cleaning, attribute transformation, and feature selection, collectively enhancing the stability and robustness of engagement rate modelling [18].

The initial dataset comprised 2,992 records and 14 attributes, requiring data cleaning to ensure accuracy, consistency, and suitability for subsequent analysis. Data cleaning involved the removal of duplicate entries, the handling of missing values in the text, musicMeta.musicName, and musicMeta.musicAuthor attributes by replacing empty captions with empty strings and unavailable music metadata with "Unknown", as well as the exclusion of invalid records such as videos with a duration of zero seconds.

In addition, numerical attributes were validated to ensure correct data types, and upload timestamps (createTimeISO) were converted from Coordinated Universal Time (UTC) to Western Indonesian Time (WIB / *Asia–Jakarta*) to maintain temporal consistency with user behaviour patterns in *Indonesia.* Irrelevant attributes were also eliminated, and a creator-level filtering strategy was applied, retaining one representative video per creator to reduce sampling bias from highly active accounts and minimise redundancy from highly correlated content within the same creator. This approach ensures that engagement patterns are learned across diverse accounts rather than being dominated by repeated content from a small number of creators, thereby improving model generalizability [19].

C. Feature Engineering

Feature engineering was performed to transform raw TikTok data into more informative feature representations to support effective classification. This stage is particularly important for algorithms such as Support Vector Machine (SVM), which are sensitive to the structure and scale of input features. In this study, feature engineering involved extracting temporal attributes from upload time, categorizing posting days into weekdays and weekends to represent audience behaviour patterns, identifying the use of trending music, and incorporating video duration as a continuous numerical feature.

To prevent data leakage between input features and the target variable, interaction-based attributes such as the number of likes, comments, shares, and saves were not included in the feature set during model training. These variables were used exclusively to compute the engagement rate label and were not provided as inputs to the Support Vector Machine (SVM) classifier. This approach ensures that the model learns audience engagement patterns from content characteristics and temporal aspects, rather than directly exploiting the variables used to construct the target label. Consequently, the methodological validity and generalizability of the proposed engagement rate classification approach are improved [20].

D. Labeling

Data labelling was performed using the Engagement Rate (ER) as the target variable for binary classification, representing the intensity of user interaction with TikTok content. Engagement rate reflects how actively audiences respond to a video through interaction behaviours, making it a relevant indicator for evaluating content performance in digital marketing contexts. In this study, ER was calculated as the ratio of total user interactions (likes, comments, shares, and saves) to total video views. Based on the resulting ER values, each video was assigned to one of two engagement categories, namely High Engagement and Low Engagement, using a threshold value of 5% as commonly adopted in prior engagement analysis studies [21]. The 5% engagement rate threshold was adopted as an operational classification criterion to ensure consistent, objective, and reproducible labelling of engagement categories across the dataset.

The 5% ER threshold was selected based on widely used global TikTok engagement benchmarks commonly reported in industry analyses, where engagement rates above 5% are generally regarded as indicative of high-performing content. Descriptive analysis of the dataset further confirms that this threshold provides balanced separation between low- and high-engagement videos, supporting its suitability for binary classification while ensuring a standardised, non-arbitrary labelling criterion.

Although interaction metrics are fundamental to defining engagement rate, these variables were used solely for label construction and were not included in the predictive feature space. This explicit separation between feature variables and the target label was adopted to mitigate information leakage and ensure a fair evaluation of the classification model.

DOI : https://doi.org/10.25139/inform.v11i1.11362

The mathematical formulation of engagement rate is presented in Equation (1). This Equation defines ER as the proportion of accumulated interaction metrics relative to the number of views, expressed as a percentage. Rather than serving solely as a numerical calculation, Equation (1) establishes a standardised labelling criterion that enables consistent differentiation between content with high and low audience engagement. The ER value produced by this formulation is subsequently utilized as the dependent variable in the SVM classification model.

$$ER = \frac{likes + comments + shares + saves}{views} \times 100 \quad (1)$$

Based on the adopted labelling scheme, the dataset exhibits moderate class imbalance, with one engagement category dominating. This characteristic can influence the classification model's learning behaviour and is therefore considered during model design and evaluation.

E. Data Normalization

During the pre-processing phase, data normalisation was applied to standardise all numerical attributes to a uniform scale before training the model. This procedure has significant value for algorithms that rely on distance calculations, such as the SVM, because variations in feature scales can skew training outcomes. The normalisation used the StandardScaler technique, which Z-scores the feature values, setting them to have a mean of zero and a standard deviation of one. Such a method diminishes the dominance of attributes possessing broader value spans and fosters a steadier, more equitable training environment, thereby enhancing the suitability of the processed data for machine learning applications focused on classifying engagement rates [22].

F. Support Vector Machine (SVM)

The categorisation of TikTok videos into engagement rate groups relied on an SVM with a linear kernel. As a supervised machine learning technique, SVM learns an optimal separating hyperplane that maximises the margin between examples from different categories, which supports robust performance in two-class classification tasks [23]. SVM was chosen for its strong capability to handle high-dimensional feature spaces while maintaining stable performance on datasets with a moderate number of samples, a condition commonly observed in social media content analysis.

Furthermore, SVM employs a transparent, weight-based decision mechanism that enables clearer interpretation of individual feature contributions. This level of interpretability is particularly important for engagement rate classification, where understanding the influence of content-related and temporal factors is as critical as achieving reliable classification performance to support informed digital marketing decisions.

The linear kernel was selected due to its computational efficiency, interpretability, and suitability for feature spaces that linear boundaries can reasonably separate. It provides an appropriate inductive bias for structured content and temporal features, where engagement patterns are better captured through additive linear relationships rather than complex non-linear transformations, thereby reducing the risk of overfitting. This kernel choice is well-suited for engagement rate classification, where content and temporal attributes are represented as structured numerical features. The SVM model was implemented using the scikit-learn library in Python, with hyperparameter tuning of the regularisation parameter C, which controls the trade-off between margin maximisation and classification error.

The overall workflow of the SVM-based classification process is illustrated in Fig. 2, which outlines the sequence from normalised feature input to model training to class prediction.

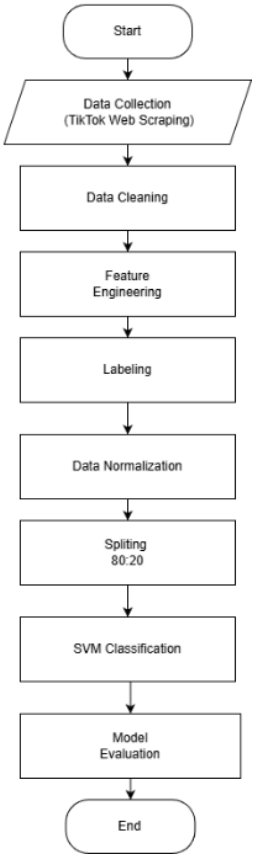

Fig.2. SVM Classification Process.

The theoretical foundation of SVM originates from the margin-based learning framework introduced by Boser, Guyon, and Vapnik, which has been shown to improve predictive reliability while reducing the risk of overfitting, particularly in high-dimensional datasets [24]. The decision function of the linear SVM model is defined in Equation (2).

$$f(x) = w^T x + b \quad (2)$$

Where the class assignment is determined based on the position of feature vectors relative to the separating hyperplane. The w variable represents the weight vector that determines the orientation of the hyperplane, $x$denotes the input feature vector, and $b$is the bias term that shifts the decision boundary. This formulation allows the classifier to assign engagement categories based on the computed decision.

DOI : https://doi.org/10.25139/inform.v11i1.11362

G. Hyperparameter Tuning

Hyperparameter tuning was performed to optimise the SVM classifier configuration for TikTok engagement rate classification. Hyperparameters are model settings specified before training that directly influence learning behaviour and decision boundary formation.

To avoid inconsistent model behaviour caused by arbitrary parameter selection, a systematic tuning strategy was applied using a grid-based search approach combined with cross-validation, which evaluates predefined combinations of hyperparameter values under consistent validation conditions. This procedure enables effective configuration of SVM parameters related to margin control and model flexibility, ensuring that the classifier is appropriately calibrated for TikTok interaction data before performance evaluation [25].

H. Model Evaluation

Model evaluation assessed the SVM classifier's effectiveness in distinguishing TikTok content across different engagement rate categories. To ensure an objective and unbiased assessment, the evaluation was performed on a test dataset that was completely separate from the training phase. Several standard evaluation metrics commonly employed in machine learning–based classification studies were utilized, including accuracy, precision, recall, and F1-score [26].

Precision and recall are complementary performance metrics that capture different aspects of classification behaviour. Precision measures the reliability of positive predictions by quantifying the proportion of correctly predicted positive instances relative to all instances the model predicts as positive. Recall, in contrast, measures the model's sensitivity by quantifying the proportion of correctly predicted positive instances relative to the total number of actual positive instances. The mathematical formulations of precision and recall are presented in Equations (3) and (4), respectively.

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

In practical classification scenarios, precision and recall are rarely interpreted independently, as improvements in one metric may lead to trade-offs in the other. Therefore, the F1-score is adopted as a unified performance metric that combines precision and recall into a harmonic mean, providing a balanced evaluation of classification effectiveness, particularly under imbalanced class distributions. The formulation of the F1-score is shown in Equation (5). Accuracy is employed as a general performance indicator that represents the proportion of correctly classified instances relative to the total number of predictions made by the model. Although accuracy provides an overall measure of classification correctness, it may not adequately reflect model performance in imbalanced classification settings, where minority classes are underrepresented. The mathematical definition of accuracy is presented in Equation (6).

III. Result and Discussion

This section presents the results of engagement rate classification using the SVM model and discusses their implications for TikTok digital marketing content. The discussion focuses on interpreting data characteristics and analytical outcomes derived from the classification process, without reiterating methodological procedures described in the previous section.

A. Data Pre-processing Overview

This subsection summarizes the outcomes of the data pre-processing stage applied to the TikTok digital marketing dataset. From an initial collection of 2,992 video records, multiple filtering procedures were applied to enhance data quality, including duplicate removal, invalid record elimination, and handling of missing values. In addition, creator-level filtering was applied by removing highly duplicated videos from the same creator, aiming to reduce redundancy while preserving representative content variations. This strategy was implemented to mitigate potential bias introduced by highly active creators and prevent the overrepresentation of specific accounts in the dataset, thereby improving sample independence and supporting more generalizable engagement-rate classification.

The overall pre-processing results are summarized in Table II, which outlines the reduction process across each cleaning stage. The substantial reduction in dataset size indicates that pre-processing played a critical role in removing noise and redundancy commonly present in raw social media data, thereby improving the reliability of subsequent classification analysis.

TABLE II
Data Cleaning Summary

Cleaning Stage	Outcome
Initial dataset	2,992 records, 14 attributes
Duplicate removal	72 records removed
Removal of zero-duration videos	70 records removed
Missing value handling	Completed
Time conversion	createTimeISO converted to *WIB* (UTC+7)
Removed attributes	authorMeta.avatar, text
Creator-level filtering	Removal of highly duplicated videos from the same creator to reduce redundancy
Final dataset	1,638 records

B. Labeling

In the pre-processing stage, each TikTok video was assigned an engagement label based on its calculated Engagement Rate (ER). A binary labelling scheme was applied with a 5% threshold: videos with ER values exceeding the threshold were categorised as High Engagement, while those below it were categorised as Low Engagement. This threshold-

DOI : https://doi.org/10.25139/inform.v11i1.11362

based labelling approach establishes a clear distinction between engagement performance levels and serves as the target variable for the classification model.

The distribution of engagement labels is presented in Table III and visualised in Fig.3. Out of 1,638 TikTok videos, 1,094 (66.79%) are categorised as High Engagement, while 544 (33.21%) are classified as Low Engagement. This distribution shows a moderate class imbalance, with the High Engagement category being the most prevalent. However, both engagement classes remain sufficiently represented, enabling the Support Vector Machine (SVM) classifier to learn meaningful and discriminative patterns from each category without excessive dominance from a single class, thereby supporting effective decision boundary formation.

TABLE III
ENGAGEMENT RATE LABEL DISTRIBUTION

Label	Category	Number of Samples	Percentage
1	High	1094	66.79%
0	Low	544	33.21%
	Total:	1638	100%

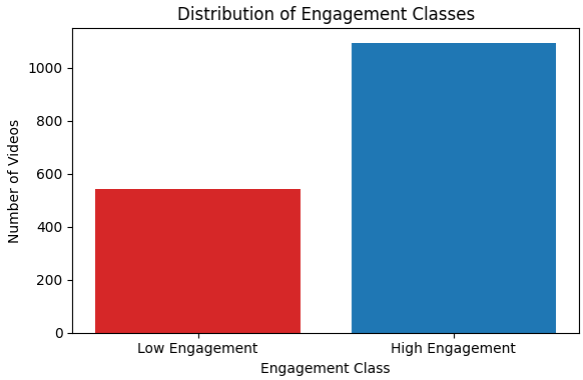

Fig.3. Distribution of High and Low Engagement Rate Labels

The label distribution presented in Table III and Fig. 3 indicates a moderate class imbalance, with High Engagement content constituting the dominant class. This condition is inherent to the labelled dataset and may influence the classification model's learning behaviour, particularly in distinguishing between engagement categories. Therefore, model performance is evaluated not only on overall accuracy but also using class-wise evaluation metrics to provide a more comprehensive assessment.

C. *Dataset Splitting*

To support objective model evaluation, the dataset was split into training and test sets at an 80:20 ratio using stratified sampling. This strategy was applied to preserve the distribution of engagement labels across both subsets, thereby reducing potential bias during model evaluation.

As summarised in Table IV, the majority of data instances were allocated to model training, while a smaller portion was reserved for testing. This configuration ensures sufficient data availability for learning while maintaining an independent dataset for performance assessment. The dataset partitioning is fully described in Table IV.

TABLE IV
DATASET SPLITTING SUMMARY

Category	Number of Samples	Percentage
Training Data (X_train)	1310	80.00%
Testing Data (X_test)	328	20.00%
Total:	1638	100%

D. *Baseline Support Vector Machine (SVM)*

The baseline Support Vector Machine (SVM) model was implemented using basic numerical features with default hyperparameter settings, without any hyperparameter tuning. The classification performance of the baseline SVM model is presented in Table V. Evaluation on the test dataset yielded an overall accuracy of 70.43%, indicating moderate classification performance. However, accuracy alone is insufficient to fully characterize model behaviour, particularly under moderately imbalanced class distributions.

A more detailed examination of the class-wise performance reveals that the baseline SVM model exhibits uneven predictive capability across engagement categories. As shown in Table V, the High Engagement class achieved a recall of 0.8037 and an F1-score of 0.7840, indicating that most high-engagement videos were correctly identified. In contrast, the Low Engagement class demonstrated weaker performance, with a recall of only 0.5046 and an F1-score of 0.5314, reflecting limited sensitivity in identifying low-engagement content. This imbalance in class-wise prediction performance is further illustrated by the confusion matrix of the baseline SVM model shown in Fig.4, which highlights a higher misclassification rate for the Low Engagement class.

TABLE V
BASELINE SVM CLASSIFICATION

Class	Precision	Recall	F1-score	Support
Low Engagement	0.5612	0.5046	0.5314	109
High Engagement	0.7652	0.8037	0.7840	219
Accuracy			0.7043	328
Macro Average	0.6632	0.6541	0.6577	328
Weighted Average	0.6974	0.7043	0.7000	328

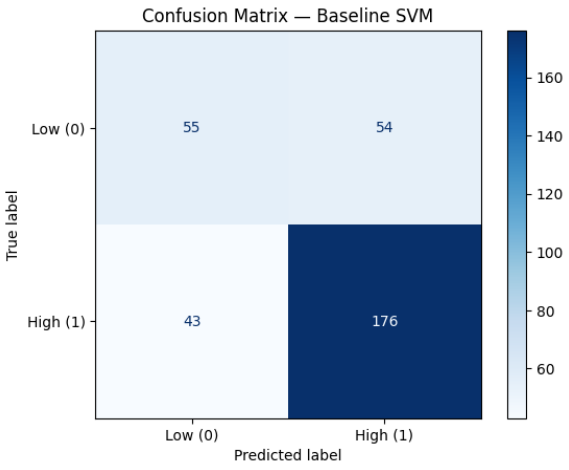

Fig.4. Confusion Matrix Baseline SVM

DOI : https://doi.org/10.25139/inform.v11i1.11362

This disparity in class-wise performance is further evidenced by the relatively low macro-average F1-score of 0.6577, compared to a higher weighted-average F1-score of 0.7000, which is influenced by the dominance of the majority class. These findings indicate that the baseline SVM model tends to favour the High Engagement category and lacks sufficient discriminatory power to identify low-engagement content consistently. Consequently, the baseline results provide a critical reference point for subsequent model optimisation, highlighting the limitations of default SVM configurations and motivating the application of systematic hyperparameter tuning and kernel selection to achieve more balanced, reliable classification performance.

E. Kernel Comparison

An evaluation was conducted to compare the effectiveness of three Support Vector Machine (SVM) kernel types: linear, polynomial, and radial basis function (RBF) using GridSearchCV with 5-fold stratified cross-validation. The comparison employed the macro-average F1-score to ensure balanced performance across engagement classes under moderate class imbalance. As shown in Fig. 5, the linear kernel achieved the highest macro-average F1-score (0.849), outperforming the RBF kernel (0.834) and the polynomial kernel (0.743). The optimal configuration for the linear kernel was obtained with *C=100*, indicating effective margin optimisation for engagement rate classification.

Although the RBF kernel is often considered more suitable for complex social media data, the superior performance of the linear kernel in this study can be attributed to the dataset's characteristics and the post-hoc classification objective. The features used, such as play count, likes, comments, shares, collections, and video duration, are aggregated numerical interaction metrics that exhibit largely monotonic and near-linear relationships with the engagement label derived from a fixed threshold. Under these conditions, a linear decision boundary is sufficient to effectively separate engagement classes, offering greater generalisation stability and a lower risk of overfitting than more flexible non-linear kernels. Based on these empirical and theoretical considerations, the linear kernel was selected as the optimal configuration for the final SVM model.

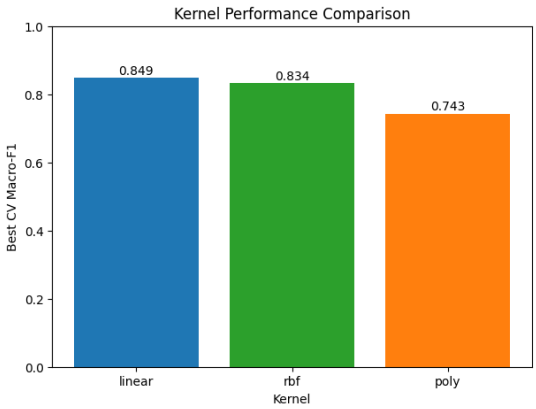

Fig.5. Comparison of SVM Kernel Performance Using 5-Fold Cross-Validation

F. Hyperparameter Tuning

The optimized SVM classifier achieved an overall accuracy of 88.41% on the testing dataset. As presented in Table VI, the Low Engagement class obtained a recall of 0.9817 and an F1-score of 0.8492, while the High Engagement class achieved a precision of 0.9892 and an F1-score of 0.9059. The corresponding confusion matrix in Fig.6 shows that 107 Low Engagement instances and 183 High Engagement instances were correctly classified.

TABLE VI
OPTIMIZED SVM CLASSIFICATION

Class	Precision	Recall	F1-score	Support
Low Engagement	0.7483	0.9817	0.8492	109
High Engagement	0.9892	0.8356	0.9059	219
Accuracy			0.8841	328
Macro Average	0.8687	0.9086	0.8776	328
Weighted Average	0.9091	0.8841	0.8871	328

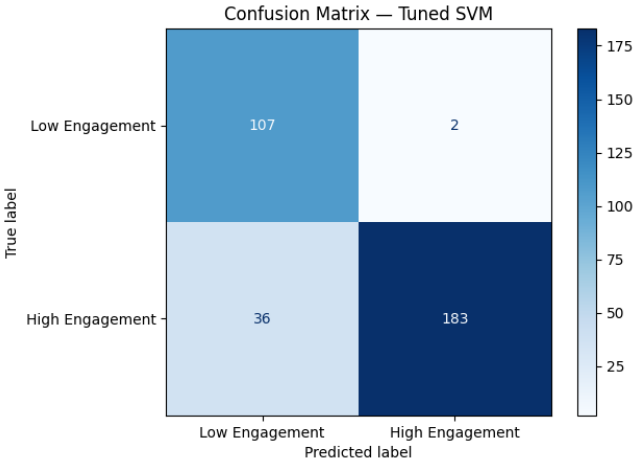

Fig.6. Confusion Matrix Tuned SVM

These results indicate that hyperparameter tuning improves the model's ability to distinguish between engagement levels compared to the baseline configuration, particularly in identifying low-engagement content. Applying class weighting improves sensitivity to the minority class, thereby reducing the bias observed in the baseline model. From a digital marketing perspective, this balanced classification supports more reliable post-hoc evaluation of content performance rather than pre-publication engagement prediction. It is important to note that the proposed SVM model is intended for post-hoc engagement categorization rather than future engagement prediction, thereby reducing the risk of overfitting claims in dynamic real-world settings.

G. Model Interpretation and Upload Recommendations

Based on the Support Vector Machine (SVM) training results with a linear kernel using non-interaction features (content and temporal attributes), the model coefficients are presented in Table VII. In a linear SVM, each feature contributes linearly to the *decision function* that defines a sample's position relative to the classification hyperplane. Consequently, the coefficient values directly indicate the

DOI : https://doi.org/10.25139/inform.v11i1.11362

direction and magnitude of each feature's influence on the probability that a content item is classified as high engagement, with positive coefficients increasing this likelihood and negative coefficients decreasing it. The absolute magnitude of each coefficient reflects the feature's relative importance in shaping the model's decision boundary.

TABLE VII
FEATURE INFLUENCE ON HIGH ENGAGEMENT PROBABILITY

Feature	Coefficient	Absolute Coefficient
videoMeta.duration	1.3834	1.3834
music_author_known	-0.1360	0.1360
music_is_original	-0.0558	0.0558
upload_hour	-0.0553	0.0553
upload_is_weekend	-0.0310	0.0310
upload_dayofweek	0.0266	0.0266
music_is_known	~0.0000	~0.0000

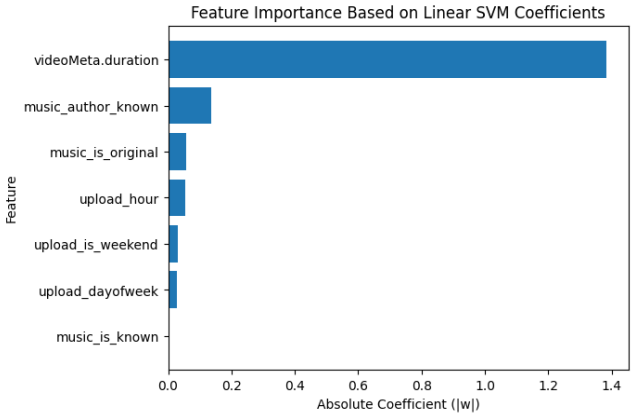

Fig.7. Feature Importance Based on Linear SVM Coefficients

As shown in Table VII and visualized in Fig.7, video duration (*videoMeta.duration*) exhibits the largest positive coefficient, indicating that duration is the most influential feature in the linear SVM decision function. Because a linear kernel is employed, feature importance can be directly inferred from the absolute values of the SVM coefficients, enabling transparent and reproducible interpretation as presented in Fig. 7. This suggests that, within the standardized feature space, longer video durations generally increase the likelihood of a content item being classified as high engagement. However, because the SVM model captures only linear relationships, this result should be interpreted as a global trend rather than a direct causal relationship. Other features, including upload timing and audio-related attributes, show smaller coefficient magnitudes in Fig. 7, indicating more moderate contributions to engagement probability than video duration.

This model-based interpretation is further supported by the descriptive analysis presented in Fig. 8, which illustrates the relationship between video duration and average engagement rate. The results indicate that videos with durations ranging approximately from 60 to 180 seconds demonstrate more stable and higher engagement performance, with several mid-range durations exhibiting notable engagement peaks. This pattern suggests that medium-length videos provide sufficient time for storytelling, explanation, or persuasive messaging, thereby enhancing audience interaction. While shorter videos remain effective, the observed trend indicates that videos in the 60–180 second range yield more consistent engagement outcomes in the analysed dataset.

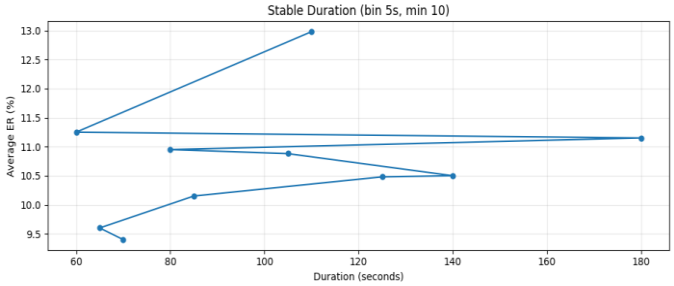

Fig.8. Average Engagement Rate by Video Duration

To complement the model-based interpretation, a descriptive analysis of engagement patterns was conducted to derive practical upload recommendations. Fig. 9 shows that several upload hours correspond to optimal time windows for content publication, likely reflecting higher user activity and content consumption.

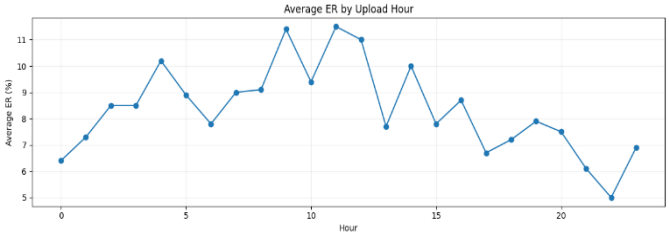

Fig.9. Average Engagement Rate by Upload Hour

In terms of upload days, Fig. 10 shows that videos published on Wednesdays, Fridays, and Saturdays achieve higher average engagement rates than those published on other days. This pattern indicates favourable audience responsiveness during midweek and the transition into the weekend, making these days strategically advantageous for digital marketing campaigns.

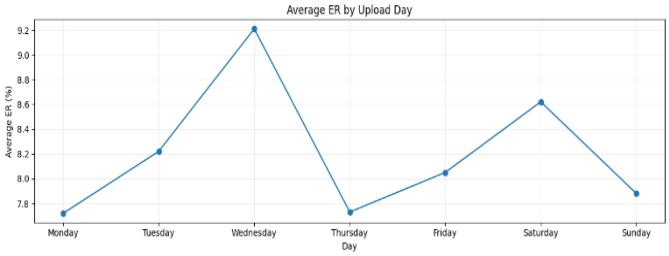

Fig.10. Average Engagement Rate by Upload Day

A comparison between weekday and weekend uploads, presented in Fig.11, reveals that weekday uploads achieve slightly higher average engagement rates than weekend uploads. Although the difference is relatively small, this finding suggests that weekday posting may offer a marginal advantage in maintaining consistent audience interaction.

Audio-related characteristics further influence engagement dynamics. As depicted in Fig. 12, videos with non-original sounds consistently achieve higher average engagement rates

DOI : https://doi.org/10.25139/inform.v11i1.11362

than those with original audio. Additionally, content using less popular or non-mainstream music outperforms videos featuring popular tracks. These findings suggest that engagement performance is not driven solely by audio popularity, but rather by the contextual relevance and alignment of music with video content and audience preferences.

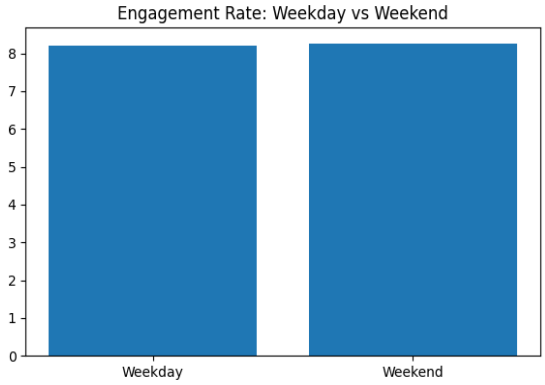

Fig.11 Engagement Rate: Weekday vs. Weekend

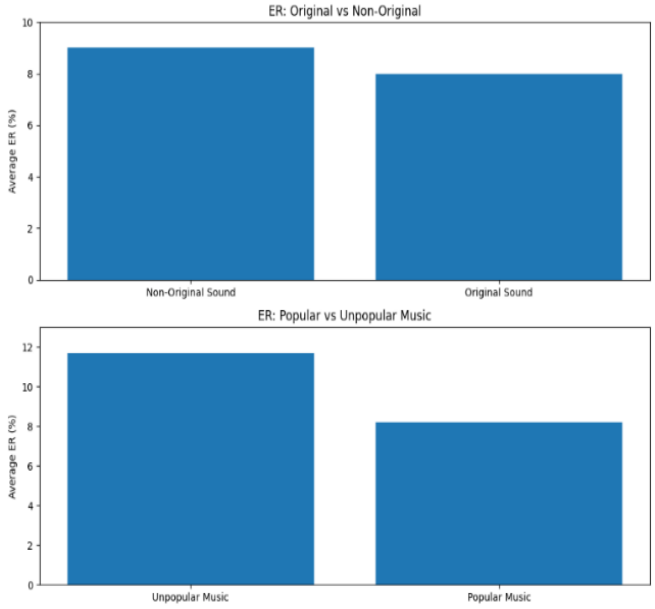

Fig.12 Average Engagement Rate by Music Type

Overall, the combined interpretation of the linear SVM model and descriptive engagement analysis provides actionable insights for optimizing TikTok digital marketing strategies. Although the features were modelled independently, the descriptive analysis suggests that certain combinations of content-related and temporal attributes, such as video duration and upload timing, may jointly influence engagement patterns. By aligning upload timing, audio selection, and content duration with the identified patterns, content creators and digital marketers can improve the likelihood of achieving higher audience engagement while maintaining methodological validity and avoiding information leakage. These recommendations are derived from post-hoc descriptive patterns and should be interpreted as strategic guidance rather than deterministic rules.

IV. Conclusion

An ER classification framework for TikTok digital marketing content was developed using a Support Vector Machine (SVM) based on content-related and temporal features. From an initial dataset of 2,992 TikTok videos, 1,638 representative samples were retained after data cleaning and creator-level filtering and evaluated using stratified sampling. The optimised SVM model demonstrated substantial performance improvement, achieving an accuracy of 88.41% compared to the baseline of 70.43%. Model interpretation indicates that video duration is the most influential factor, followed by temporal attributes and audio characteristics. Descriptive analysis further highlights that higher engagement is associated with medium-length videos (approximately 60–180 seconds), uploads during late morning hours (09:00–12:00), midweek-to-early-weekend posting, and the use of non-original, less popular music.

The proposed framework is intended for engagement classification rather than prediction and serves as an interpretable decision-support tool for digital marketing optimization. In addition, its feature-based methodology enables potential adaptation to other short-form video platforms, such as Instagram Reels and YouTube Shorts, by adjusting platform-specific features and interaction mechanisms, while acknowledging that platform-specific algorithms and user behaviour differences should be considered in future studies. The proposed model can be further developed by incorporating hashtag usage strategies, caption length as a linguistic feature, and audio novelty to enrich engagement analysis and provide strategic recommendations for digital marketing practices.

References

[1] A. A. Irwanda *et al.*, "Analisis Engagement Rate Pada Instagram Universitas Lancang Kuning," *Zo. J. Sist. Inf.*, vol. 6, no. 2, pp. 391–399, 2024, doi: https://doi.org/10.31849/zn.v6i2.17904.

[2] D. Wiryany, S. Natasha, and R. Kurniawan, "Komunikasi Terhadap Perubahan Sistem," *J. Nomosleca*, vol. 8, no. November, pp. 242–252, 2022, doi: https://doi.org/10.26905/nomosleca.v8i2.8821.

[3] N. Sapina, A. Nanda, and M. A. Arifin, "Analysis of Factors that Influence Video Engagement on the TikTok Platform Using the Multiple Linear Regression Algorithm Analisis Faktor-Faktor yang Mempengaruhi Engagement Video di Platform TikTok Menggunakan Multiple Linear Regression," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 5, no. July, pp. 875–885, 2025, doi: https://doi.org/10.57152/malcom.v5i3.1987.

[4] Rahmi Rosita and Evalina Darlin, "Pengaruh Kualitas Konten Tik Tok Terhadap Customer Engagement Pada Customer Queensha," *J. Lentera Bisnis*, vol. 13, pp. 1061–1071, 2024, doi: 10.34127/jrlab.v13i2.1129.

[5] C. A. Krisdanu and K. A. Sumantri, "TikTok sebagai Media Pemasaran Digital di Indonesia Cheryl Arshiefa Krisdanu 1 , Kiranastari Asoka Sumantri 2 12," *J. Lensa Mutiara Komun.*, vol. 7, no. 2, pp. 24–36, 2023, doi: https://doi.org/10.51544/jlmk.v7i2.4173.

[6] T. Tatasari, S. Purnomo, and A. K. Dewa, "Social Sciences Journal (SSJ) Pemanfaatan Konten Digital Berbasis Video Pendek untuk Meningkatkan Engagement pada UMKM Makanan di Media Sosial," *Soc. Sci. J.*, vol. 3, no. 2, pp. 20–32, 2025.

[7] V. V. Madnure and D. P. A. Kadam, "Predictive Modeling of User Engagement Patterns On Social Media Using Data Mining Approaches," *Int. J. Sci. Technol.*, vol. 16, no. 4, pp. 1–11, 2025, doi: 10.71097/IJSAT.v16.i4.8876.

[8] S. Zannettou, O. Nemes-nemeth, A. Goetzen, K. P. Gummadi, and E. M.

DOI : https://doi.org/10.25139/inform.v11i1.11362

Redmiles, "Analyzing User Engagement with TikTok' s Short Format Video Recommendations using Data Donations", doi: 10.1145/3613904.3642433.

[9] N. Lau, K. Srinakarin, H. Aalfs, X. Zhao, and T. M. Palermo, "TikTok and teen mental health : an analysis of user-generated content and engagement," *J. Pediatr. Psychol.*, vol. 50, no. July 2024, pp. 63–75, 2025, doi: https://doi.org/10.1093/jpepsy/jsae039.

[10] M. B. Younes, "Assisting Drivers at Stop Signs in a Connected Vehicle Environment," *Futur. Internet*, vol. 15, no. 7, p. 238, 2023, doi: https://doi.org/ 10.3390/fi1507023.

[11] S. Rahmatullah and Sriani, "Identification of Social Media Addiction Levels on " TikTok " Among Students Using Mamdani Fuzzy Logic," *Inf. J. Ilm. Bid. Teknol. Inf. dan Komun.*, vol. 10, no. 1, pp. 8–15, 2025, doi: https://doi.org/10.25139/inform.v10i1.8447.

[12] F. H. Nasution, R. A. Setyanti, and Y. Siregar, "Analisis Faktor Faktor Yang Mempengaruhi Viralnya Konten Tiktok Dengan Pendekatan Statistik," *J. Artif. Intell. Digit. Bus.*, vol. 4, no. 2, pp. 2441–2445, 2025, doi: https://doi.org/ 10.31004/riggs.v4i2.875.

[13] M. H. Arfian *et al*., "Analisis Sentimen Pada Media Sosial Menggunakan Metode Support Vector Machine," *J. Ilmu Tek. dan Komput.*, vol. 09, no. 01, pp. 1–6, 2025, doi: http://dx.doi.org/10.22441/jitkom.v9i1.001.

[14] I. R. Ainunnisa and S. Sulastri, "Analisis Sentimen Aplikasi Tiktok dengan Metode Support Vector Machine (SVM), Logistic Regression dan Naïve Bayes," *J. Teknol. Sist. Inf. dan Apl.*, vol. 6, no. 3, pp. 423–430, 2023, doi: 10.32493/jtsi.v6i3.31076.

[15] T. Alifah, D. Wahdiyati, and N. Rahman, "Analisis Engagement Rate Pada Konten Video di Akun Tiktok Grup Idol Virtual Plave @ plave official," *J. Komun. Nusant.*, vol. 7, no. 1, pp. 1–15, 2025, doi: https://doi.org/10.33366/jkn.v7i1.226.

[16] M. Furqan, R. Ichsan, and H. Hasibuan, "Recognition Of Calligraphy Writing Patterns Using The Zernike Moment Method And Support Vector Machine," *J. Sci. Soc. Res.*, vol. 4307, no. 4, pp. 2139–2146, 2024, doi: https://doi.org/10.54314/jssr.v7i4.2362.

[17] A. S. Chan, M. Husna, P. P. Putra, A. Info, S. Analysis, and T. Classification, "A Performance Enhancement Strategy for Sentiment Classification Models On Political Social Media Using Hyperparameter Tuning And Boosting," *Int. J. Adv. Data Inf. Syst.*, vol. 6, no. 3, pp. 763–772, 2025, doi: 10.59395/ijadis.v6i3.1455.

[18] C. Maulida, T. Yunanda, M. Hanafi, W. Mega, and P. Dhuhita, "Sentiment Analysis on TikTok Shop Reviews Using Long Short-Term Memory Method to Find Business Opportunity," *Inf. J. Ilm. Bid. Teknol. Inf. dan Komunikas*, vol. 9, no. 1, pp. 1–7, 2024, doi: https://doi.org/10.25139/inform.v9i1.6524.

[19] A. Agung, A. Daniswara, and I. K. D. Nuryana, "Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru," *JINACS (Journal Informatics Comput. Sci.*, vol. 05, no. 01, pp. 97–100, 2023, doi: https://doi.org/10.26740/jinacs.v5n01.p97-100.

[20] N. Putri *et al*., "Penerapan Feature Engineering Dan Hyperparameter Tuning Untuk Meningkatkan Akurasi Model Random Forest Pada Application Of Feature Engineering And Hyperparameter Tuning To Improve The Accuracy Of Random Forest Models On Credit Risk," *J. JTIK (Jurnal Teknol. Inf. dan Komunikasi)*, vol. 12, no. 2, pp. 251–262, 2025, doi: 10.25126/jtiik.2025128472.

[21] V. Padiya, D. Shah, T. Dave, and R. Karani, *Classification of Instagram Users and Prediction of Engagement Rates Using Machine Learning*. Springer Nature Singapore, 2025. doi: 10.1007/978-981-96-4139-0.

[22] I. Permana and F. Nur Salisah, "The Effect of Data Normalization on the Performance of the Classification Results of the Backpropagation Algorithm Pengaruh Normalisasi Data Terhadap Performa Hasil Klasifikasi Algoritma Backpropagation," *IJIRSE Indones. J. Inform. Res. Softw. Eng. J.*, vol. 2, no. 1, pp. 67–72, 2022, doi: https://doi.org/10.57152/ijirse.v2i1.311.

[23] J. Anggraini and D. Alita, "Implementasi Metode SVM Pada Sentimen Analisis Terhadap Pemilihan Presiden (Pilpres) 2024 Di Twitter," *J. Inform. J. Pengemb. IT*, vol. 9, no. 2, pp. 102–111, 2024, doi: 10.30591/jpit.v9i2.6560.

[24] Sriani, Sulindawaty, and Y. Rizky, "Tembakau Menggunakan Glcm (Gray Level Co-Occurrence Matrix) Dan Svm (Support Vector Machine)," *JITET (Jurnal Inform. dan Tek. Elektro Ter.*, vol. 12, no. 3, pp. 3342–3349, 2024, doi: http://dx.doi.org/10.23960/jitet.v12i3.4599.

[25] W. Nugraha and A. Sasongko, "Hyperparameter Tuning pada Algoritma Klasifikasi dengan Grid Search Hyperparameter Tuning on Classification Algorithm with Grid Search," *Sist. J. Sist. Inf.*, vol. 11, no. 2, pp. 391–401, 2022, doi: https://doi.org/10.32520/stmsi.v11i2.1750.

[26] F. Farasalsabila, V. B. Lestari, D. D. N. Cahyo, T. Lestari, F. Rusdi, and A. Islami, "Sentiment Analysis for IMDb Movie Review Using Support Vector Machine (SVM) Method," *Inf. J. Ilm. Bid. Teknol. Inf. dan Komunikas*, vol. 8, no. 2, 2023, doi: https://doi.org/10.25139/inform.v8i2.5700.

DOI : https://doi.org/10.25139/inform.v11i1.11362